THE RELIABILITY OF SINGLE MEASUREMENTS WITH STANDARD TESTS[1]

S. A. COURTIS
Liggett School, Detroit, Mich.

The question of the reliability of a single measure of a child's ability to write the answers to the addition combinations seems to the writer better shown by repeated measurements at intervals over a long period of time than by the creation of an artificial practice series. Accordingly he presents in Fig. 8 the scores of certain members of an eighth-grade class under his control. The period of time represented is three school years. The test the first year differed from that shown in Fig. 1 above in that it contained no zero combinations; that is, the test was slightly more difficult. For the other two years one or another of the three editions of Test 1 were used. Variations in external conditions were reduced to a minimum through mechanical timing and uniform procedure.

It will be noted that the curve based upon the class averages show marked fluctuations. Through October, November, and December, 1909, the curve is level. During this period the class was drilling on the combinations in the other operations, and the work did not influence the scores in addition; did not "transfer" in other words. Accordingly this is a good period to study the fluctuations in the scores of individuals. During the remainder of this school year there was direct practice on the addition combinations and the scores show corresponding increases. The first week in October, 1910, the addition test was given each day of the week, and during the following eight weeks, once each week, thus forming a direct practice series. As a whole the 1910–11 curve of progress after the first rise is fairly level during this period, with gain as before during the two remaining thirds of the year. The last year the tests were given but three times, little attention was

[1] Continued from Vol. XIII, No. 7, March, 1913.

paid to direct work on the tables, and most of the gain was made the first third.

Tabulation of the November to December differences (19 cases) gave an average variation of 6 points and a maximum of 13; of the December to January differences (Christmas vacation) an

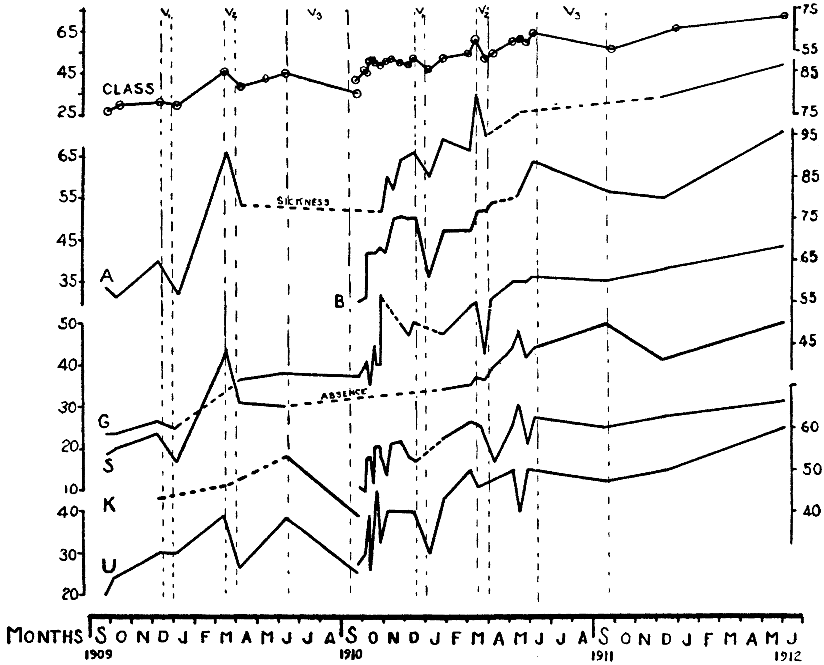

FIG. 8.—Curves based upon the average scores in Test 1, addition combinations, made by a class, and upon the individual scores of certain members in the class. Time scale, months from September, 1909, to June, 1912. Small circles upon class curve indicate the dates at which tests were given. V_1, V_2, V_3 represent the Christmas, spring, and summer vacations respectively. Scales at the right and left give number of answers written per minute. Note that a unit space in the class scale has twice the value of the unit in the other scales. For explanation of the curves see text.

average variation of 5 points. The average difference from one score to the next was for the practice week (97 cases), 4.7 points; 63 per cent of the differences were 5 or under; 25 per cent from 6 to 10 points, and 11 per cent from 10 to 17 points. For the eight-week period (124 cases) the average variation was 3.9 points;

74 per cent of these differences were less than 5 points; 23 per cent from 6 to 10 points, and 5 per cent from 10 to 17 points.

These figures agree very well with the results of the article under discussion and show that in general a single measurement of an individual will yield a value that will not differ more than 5 points from his true ability at the time. For two-thirds of the children the differences will be 5 or less. About one child in ten will have a markedly unreliable score, the amount of the deviation ranging from 10 to 30 points.

So far, the results fully confirm the data from the other study and with such general statements the writer has no quarrel. His contention, however, is that the fact of the great range in the nature and amount of individual variation makes such statements of little value unless rightly interpreted. The child is a plastic, developing organism. Its performance in any test is determined by a large number of factors. The nature of some of these factors, and the degree of the effects of all are entirely unknown. If the child's mind is in stable equilibrium, a single measurement (barring external accidents) will yield a reliable result, and a series of measurements will show but small chance variations from the score of the first test. If, however, the child's mind is in unstable equilibrium so that it is easily modified by experience, any one score may differ markedly from any number of previous or later measurements that may be made, and a practice series may yield results that it will be as foolish to combine into a single measure of ability as it is to express the sum of three apples and two potatoes by a single figure with the name of either. In one sense each and every measurement is absolutely unreliable in that it is impossible to predict, on the basis of any number of past measurements, what score an individual will make in the next test, owing to possible changes made in the mind of the individual by that past work. In another sense each independent measurement is also absolutely reliable; for it records the actual achievement of the individual at the time and under the conditions of the test. In other words, the writer believes each mental measurement should be treated only as symptom of internal conditions. An unusual score is a sign that a cause or causes are at work which must be evaluated

before the score can be rightly interpreted, but the score itself is a true measure of conditions at that instant, and, if the conditions continue, the score of a second test will agree with the first within two or three points. But no general rule can be made that will cover all the facts of individual variation, and the interpretation of one or more measurements of an individual must remain a matter of individual judgment.

A consideration of the individual curves given in Fig. 8 will make these points clear. Individual A the first week of school made a score of 33 and two weeks later of 31. In December before the Christmas vacation her score was 40, after vacation 32. The class average during the same period shows little variation. One may interpret these differences in these scores as chance variation, or one may see from the rest of the curve that this individual is easily affected by her experiences. When the class is gaining, she gains a great deal; during vacation she loses consistently. It is probable, therefore, that the change from 31 to 40 in December represents a real internal change, a "transfer" from the regular work in arithmetic. Her actual score at the end of any practice series will be unreliable by 10 or more points, since, during any period of disuse, her ability to add decreases by that amount. On the other hand, it is interesting to note that a large part of the gain made during the midyear term, 1909–10, was permanent, as her scores after vacation and after a long and serious illness agree. This individual makes a fine showing under drill, but the gain is largely specific and misleading.

Individual B, on the other hand, is of quite a different type and her curve is given as representing that of a girl of unusual ability. The opening week of school her score was 55, the Monday of the following week it was 56. On Tuesday it jumped without warning to 67, and on the remaining days of the week her scores were 67, 67, 67. The following Monday it was still 67, and in the three tests during the following month her scores were 67, 68, 67, a series of seven scores practically without variation. The second week in November her score again increased suddenly, this time to 75, and from this time on to April her scores were 75, 76, 75, 75, 61, 72, 72, 77, 77, 79. The low score, 61, marks the effect of the

Christmas vacation, but the loss was quickly made up. The remaining scores for May and June were 81, 86, 89, and for the next year 82, 80, 96. Such constancy and such sudden gains are by no means unusual. The writer can show many cases where the same score has been repeated week after week for five or six weeks at a time. It would seem, too, that in many children there are "psychological moments" when a little practice will produce marked gains which are held forever after, while even a large amount of practice at other times produces little or no apparent effect. Whether this is merely the much-discussed "plateau effect," or whether it reflects certain inner physiological changes, the writer does not know.

The other curves were chosen to illustrate various types of children. G is that of a stable individual of average ability; S shows the effect of absence and travel (a little tutoring was done on the trip but no school was attended during this period); K, a variable individual of more than average ability; U, of the member of the class weakest in addition. U gained steadily during the first two-thirds of the year. The March (1910) vacation and the summer vacation following both show large losses. The practice series, 1910, was marked by violent fluctuations in score, while for the succeeding five weeks a perfectly constant score of 40 was made each week. It should be noted that 40 is the level nearly reached at the two previous high points of the curve.

One cannot follow the score of many individuals in successive tests of addition, subtraction, multiplication, and division through several years without realizing that the number of factors determining any one score is very large, and that the question of reliability is not as simple a question as it might seem. It is certain that the possible unreliability of any single measurement must always be kept in mind; it is equally certain, also, that nine times out of ten a single test reflects the true conditions accurately. It is granted that a series of twenty-five tests will reveal additional facts in regard to an individual, but it is also to be remarked that it may take many other related tests and much experimentation to make the meaning of the additional facts clear. As a practical

expedient, twenty-five tests of each individual in the hundreds of abilities that must eventually be measured if there is to be efficient classroom teaching is out of the question. The writer is sorry the conclusions of the authors should have taken this form, and hopes that no teacher or superintendent just awakening to the value and possibilities of measurement and standardization will be discouraged by the statement. For the purpose of the study was based upon a misconception of the purpose of the tests, and the methods used are applicable only to more stable mental conditions than are found in the minds of growing children.

In this connection it is necessary to remind the reader that Test 1 is but one of a series of eight related tests and that, as used by the writer, the scores in all these tests are interpreted together. To quote from the folder of instructions describing the use of the comparative graph sheet, "The general plan is, therefore, to interpret any curve in the light of all the knowledge of the individual that can be obtained, and to check conclusions reached by further tests at the first opportunity."

Thus in Fig. 9 are given the curves of two seventh-grade individuals of extreme types. In the interpretations of such curves the scores made in Tests 7 and 8 should be considered first. For Test 7 the standard scores for the seventh grade are 13 examples attempted and 8 right. Individual A's score is unsatisfactory, as she has attempted 14 examples and has but 4 of the 14 right. Individual B's score, on the other hand, represents exceptional ability, for although she attempted but 11 examples, 10 of these were right, indicating unusual accuracy. Now it should be evident at once that since B actually has the ability to add, subtract, multiply, and divide in Test 7, it is of little consequence whether he knows his tables or not. The fact that in the addition combinations his score is that of a fourth-grade child has no significance whatever. He can actually add in long-column addition at a higher rate of speed than he can write the answers to the single combinations, and on the basis of much experience in this type of work, the writer is able to state that any attempt to increase his score in Tests 1 to 4 by practice is likely to act adversely upon Test 7. Yet it is certain that the low score of 35 in addition is not an "unreliable"

score; for it is confirmed by the scores in the other speed tests. Even the higher score in division is probably not due to chance variation, but to the greater ability of the individual in this process.

The unsatisfactory scores of individual A in Test 7, however, are not due to "tables" either; for her scores are two grades above standard in all except division, and are double those of individual B. Again these scores are reliable, for the four scores agree closely.

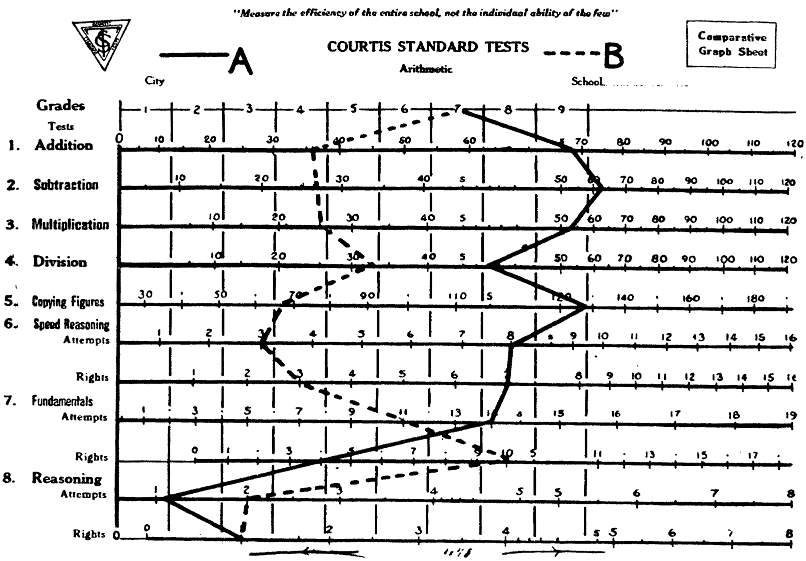

FIG. 9.—Curves of two individuals of extreme types. A (solid line) knows her tables very well, but works inaccurately in Test 7 (abstract examples in the four operations), getting but four examples right out of the 14 attempted. B (broken line) has only 4th-grade scores in the tables, but is able to work ten examples out of eleven correctly. Neither A nor B would profit by drill on the tables.

The apparent drop in division is due to the character of the scale at this point where the maximum of the development curve occurs, and the real difference is so small that it falls within the range of chance errors (5). A has been overdrilled on the tables. It should be evident that no amount of further drill on the tables could benefit A in the slightest, *unless* her past drill has been wholly written, so that the score is an indication of the specific ability to write the answers and not of ability to make ready response.

An oral test could be used to answer this question. It is probable, however, that A's defect is to be found in some of the other abilities involved in the longer computations, such as ability to "borrow and carry," to copy figures accurately, etc. Whatever the cause, it must be remedied by special work at the weak point, and no other remedial work will be efficient. The actual cause could easily be determined from an analysis of the mistakes made.

It should be noticed that in the interpretation of such departures from standard, balance of scores is more important than absolute size of score, and that so small a difference as ten points is without great significance. The average yearly progress is about seven points. Accepting the author's own figures, the scores of eight children out of nine will fall within ten points or within a grade and a half. More than two-thirds of a class will differ less than a grade. How accurately a single measurement thus places an individual on the scale of status and how great the need for measurement in teaching is shown by the statistics for the range of individual variation within the grades under present inefficient conditions.

In Fig. 10 is given the distribution of 7,625 eighth-grade scores in Test 1, also for comparison, the distributions of a number of groups of children in single classes from various cities, and of 118 eighth-grade class averages. It will be seen that the range is very great and fairly constant from city to city. The eighth-grade average score in addition is about the same everywhere.[1] As long as the differences between individuals in the same class amounts to as much as the entire average progress during the whole school life, a test that will enable a teacher to place a child on the scale of status as accurately as within ten points and on the basis of a single measurement, must be of very great value to the intelligent teacher. As long as present conditions continue, the need for the refinements of measures derived from practice series of twenty-five or more measurements of each individual is not apparent.

Such curves as those of B above raise a very important question, "What is the ideal form of the individual development curve?"

[1] The larger cities tend to emphasize the abstract at the expense of the reasoning work, and the average scores in Test 1 are from ten to fifteen points higher, with a corresponding range of individual variation.

In Fig. 11 is given the actual development curve for knowledge of the fundamental combinations in addition based upon the standard scores derived from the measure of many thousands of children (light solid line). A form in accordance with the suggestion made

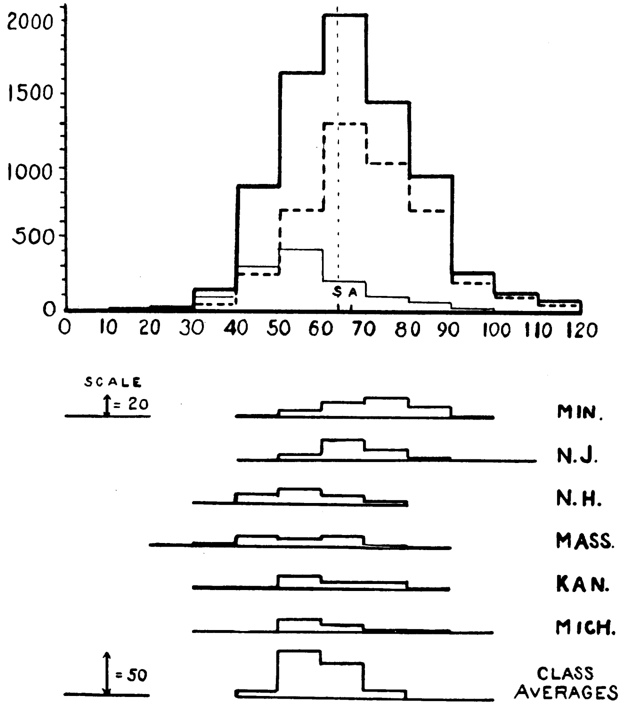

Fig. 10.—Upper part of figure, distribution of scores of 7,625 eighth-grade children in Test 1. Collected from schools in many states. Vertical scale = number of children making each score. Horizontal scale = number of answers written per minute. The range of the scores is from 15 answers per minute to 115 per minute. A = average score of the entire group; S = standard eighth-grade score.

Middle figures = distributions of individual scores in typical classes from various states.

Lower figure = distribution of 118 eighth-grade class averages; 80 per cent fall within a range of 20 points.

by the writer above—that the growth occurs suddenly and at certain periods only—is indicated by the broken line. The dotted line is a variation of this same idea; i.e., that the child should learn his table early and all at once. Slowly gathering evidence has led

the writer to suspect, however, that the ideal form for efficient teaching is that represented by the heavy line; that is, that under ideal conditions a child would acquire through concrete experience and oral drill a working knowledge of the simpler combinations. He would put this knowledge to immediate use in working abstract examples and perfect his knowledge by repeated use. But as there would be no drill on the separate combinations, before long higher habits of grouping, of unconscious short cuts, etc., would begin to form and ability to recall the separate combinations would

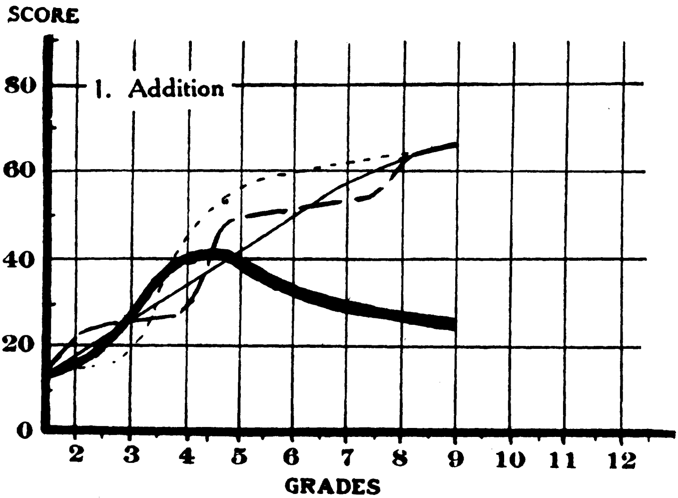

FIG. 11.—Various types of development curves for the ability measured by Test 1. Light, solid line = actual curve derived from measurement of many children. Heavy line = possible ideal form. For meaning of other lines see text.

decline. Whether this is the true explanation or not, it affords a working basis for valuable experimental work. At least two things are certain: (1) many children able to do eighth-grade work in Test 7 with fourth-grade knowledge of the tables are found, particularly among those who are exceptionally accurate: (2) a very great deal of useless, ineffective drill work on the tables is done in our schools. In Fig. 10, 762 or 10 per cent of the whole group have scores higher than the standard by 20 or more points, although very few of these really use the tables at anything like the standard rate.

One last point remains to be discussed; the question as to the correlation that exists between ability in Test 1 and ability in column addition. Fortunately in his attempt to derive units of measurement, the writer has had occasion to give a series of tests designed to measure the relative difficulty of various addition examples and that data is presented herewith.

In Table V is given a sample example from each of the series of tests. Tests A and H were the speed test shown in Fig. 1. Test B was composed of single combinations—practically the test shown in Fig. 1 with different spacing and the zero and 1 com-

TABLE V

SAMPLE EXAMPLES FROM A SERIES OF TESTS IN COLUMN ADDITION. EACH TEST WAS COMPOSED WHOLLY OF EXAMPLES OF ONE TYPE, AND WAS OF SUCH LENGTH THAT NO ONE FINISHED IN THE TIME ALLOWED

A	C	E	F	G	I	J	K	L
8	5			7				
3	6			5				
	2			8				
	7			9				
(B)			4	4	8	3	4	8
			7	3	3	49	557	9,659
7	(D)	9	2	9	4	66	892	3,778
4		3	2	7	4	75	347	9,484
	7	6	4	9	7	32	562	5,247
(H)	6	4	9	7	8	96	738	8,470
	2	8	2	8	7	85	658	7,966
9	7	2	5	5	6	64	273	6,323
2	4	3	2	2	2	59	797	3,277

binations omitted. Test C consisted of examples containing four figures, thus requiring three additions except as a person was able to "see" the sum of the four at once. Each of the other tests contained one type of examples, the examples increasing in length up to columns of 13 figures (12 additions). The examples in Tests I, J, K, L are all composed of columns of 8 additions each but increase from single columns up to examples of four columns. Such a series unquestionably measures the "ability to add."

The tests (except the speed tests) were mimeographed and given in regular class time by the class teacher. The time allowances varied from one minute for the simpler tests to three minutes

for the more difficult. Mechanical timing was used. Tests A to H were given on one day in rapid succession, from one to two minutes' rest being allowed between each trial. The remaining tests were given some weeks later. The scores were expressed as number of additions made per minute. The mistakes were scored as number of "columns wrong." The individual results are given in Table VI, also the averages for the class. Individuals A to U are eighth-grade girls and the letters represent the same individuals as in Fig. 8. Individuals X, Y, Z are teachers tested at the time the class was tested and also a year later for the purposes of this article. Y is the writer himself. Z is his assistant in charge of his statistical work, a bookkeeper with a business-college training and a year and a half of practical experience, an exceedingly rapid and efficient worker.

Even a casual inspection of the table will show that the correlation between ability in Test 1 and ability "to add" is no simple thing and like most other such relations between mental tests is wholly an individual matter. The results cannot properly be expressed in a single coefficient. For instance, individual A who stands highest in the speed Tests A, B, and H—the single combinations—falls to a rank of fifteenth when there are even three additions in a column and varies from eleventh to nineteenth in succeeding tests. B, on the other hand, is second in the speed tests and first in all other tests. At the other extreme, individual Q was seventeenth in the speed tests but is second highest in the four most difficult tests, and individual T was twentieth in the speed tests and was fifth in the last test. Between these extremes every degree of correlation is found. The correlation coefficient based upon the relative ranks of the 21 individuals in the first and last tests is $r = +0.43$.

The essential facts of the table and additional data derived from the same tests in other grades are presented graphically in Fig. 12. Test H is placed next to Test A in order that comparison may be easily made. A and H, it should be remembered, are two trials with the test shown in Fig. 1, one trial before and one after the first five tests in adding columns; that is, before and after eleven minutes' actual work in column addition. Every grade

but the sixth grade shows an increase in score and in the eighth grade the practice effect is comparable to previous results, 75 per

TABLE VI

COLUMN ADDITION

Scores in Number of Additions per Minute of 21 Eighth-Grade Girls

Test No.	A	B	C	D	E	F	G	H	I	J	K	L
A.......	85	60	33	26	23	24	22	86	36	30	25	25
B.......	77	54	50	52	54	50	42	82	70	60	58	56
C.......	77	50	30	36	33	24	24	75	36	28	32	29
D.......	73	53	48	47	40	48	39	80	46	44	39	40
E.......	72	41	37	41	45	44	42	88	44	36	38	28
F.......	71	37	33	40	30	34	28	72	36	17	30	29
G.......	69	43	36	37	31	31	28	72	36	38	29	24
H.......	69	42	37	30	29	22	22	67	40	28	26	26
I.......	65	41	36	50	24	24	24	67	28	28	26	24
J.......	65	42	33	37	31	20	20	68	28	28	33	21
K.......	65	39	32	31	24	34	30	62	40	40	38	32
L.......	64	43	39	38	38	32	34	67	40	36	39	40
M.......	62	40	42	34	34	34	32	60	44	36	33	29
N.......	62	47	40	47	43	34	37	64	36	26	34	27
O.......	62	40	36	35	37	30	26	64	44	33	29	26
P.......	55	39	29	19	17	18	14	62	36	20	22	18
Q.......	55	40	42	46	42	46	36	56	49	49	45	43
R.......	52	41	30	30	23	24	28	58	32	26	23	21
S.......	52	24	27	22	21	28	27	47	24	20	30	24
T.......	48	31	38	37	37	42	42	52	44	36	29	32
U.......	39	30	32	29	23	20	26	49	36	22	26	26
Average .	63	41	36	35	31	33	29	66	39	32	33	30
W^1......	84	63	62	76	60	58	50	93	..	..	..	..
W^2......	87	67	66	78	71	64	58	86	62	52	59	..
X^1.......	..	65	69	64	70	66	58	..	..	..	..	..
X^2.......	70	65	63	68	66	58	53	85	64	52	55	..
Y^1.......	..	70	72	67	72	69	69	..	..	..	..	..
Y^2.......	87	71	77	71	76	79	76	93	67	50	47	62
Y^3.......	85	67	78	70	68	65	60	90	70	51	55	58
Z^1.......	105	87	114	120	113	112	108	110	116	89	101	111
Z^2.......	112	90	120	116	108	120	111	113	128	96	99	115
Average for W, X, Y.......	83	69	70	71	69	66	61	89	66	51	54	..

Note.—W, X, and Y are teachers, Z a bookkeeper.

cent of the children varying five or less points, 19 per cent between 6 and 10, 5 per cent between 11 and 16. A feature of the results is that 26 per cent of the fifth-grade differences, 57 per cent of the

sixth-grade, 18 per cent of the seventh-grade, and 19 per cent of the eighth-grade differences were negative. This, together with the decline in the curves, would seem to indicate the effect of a fatigue factor. In this connection the recovery on Test I should be noted. Test I was of course the same as Test F, but in the case of the children was given first of the Tests I, J, K, and L a few weeks

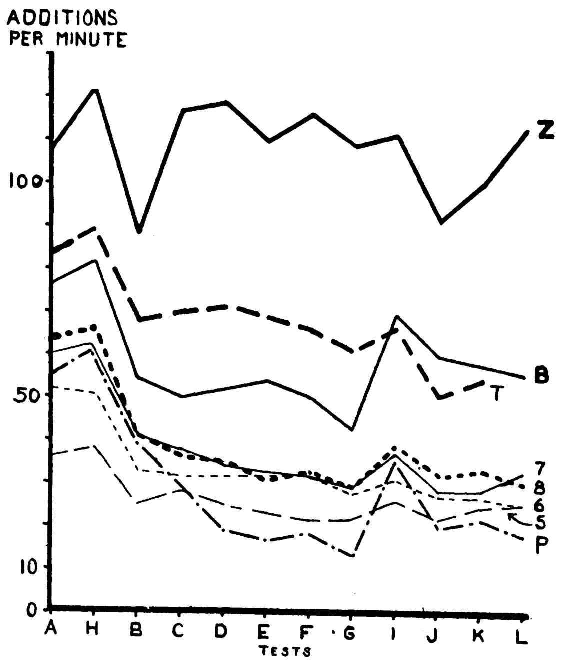

FIG. 12.—Graph of relative scores in tests of column addition. Vertical scale = number of additions per minute. Horizontal scale = tests corresponding to examples shown in Table V. Z = curve of bookkeeper; T = average scores of three mathematics teachers; B = best individual in the class; P = poorest individual. 7, 8, 6, 5 = curves of corresponding grade averages.

after the previous series. For individual Z and for the teachers (T), however, the two series were given one after the other on the same day.

In considering the results, the difference between Tests A, H, and B should be noted first. B is slightly harder than A and H because the zero and one combinations were omitted, but the real cause of the difference in score is the greater distance the hand must

travel in writing the answers. B was constructed to conform to the spacing in the succeeding tests. In Tests A and H the combinations were printed twenty on a line while in B there are but ten on a line. Even Z who is able to think the answers at a very high rate of speed cannot overcome the physical handicap of greater number of figures to be written and the greater space to be covered. The differences between the scores for A and B and between B and the other tests are quite constant. Therefore in general the individual who has a high score in the speed tests will show a corresponding score in "adding."

The general statement of the correlation between the two abilities will not be true of individuals as was shown above. In the figure, B is the curve of the best individual in the class and P of the poorest, and these two differ more in Tests E and F than they do in Tests A and H, that is, the difference between their ability in column addition is greater than the difference shown by their scores in the combinations. This does not mean, however, that the tests of the combinations do not measure abilities used in column addition; for even on the basis of the analysis of the abilities involved in column addition given by the authors, readiness of association is but one of several factors and the lack of correlation may be due to a defect in one of the other factors. The writer would add two factors to those given by the authors of the study being discussed, (1) attention span, (2) fatigue, and in this case the extreme drop of the curve of P is probably due to the effects of fatigue.

The relative level of the curves of the different grades is very significant. The seventh grade has overtaken the eighth and the sixth the seventh. The writer of course believes this due to changes in methods that have been made based upon the results of the testing work and expects to produce very much greater changes through knowledge and control of the various factors involved. It is impossible to teach efficiently a child to "add" when neither the teacher nor the children have more than such a vague general aim before them. At the same time the separation between the eighth-grade average and the curve of the teacher's score and between the teacher's and professional ability in addition show that

conditions are far from satisfactory. It is a relatively simple matter to develop "readiness of response," but to teach column addition is quite a different matter.

Some of the individual results from such tests, however, throw light upon the factors determining a score. In Fig. 13 are given a

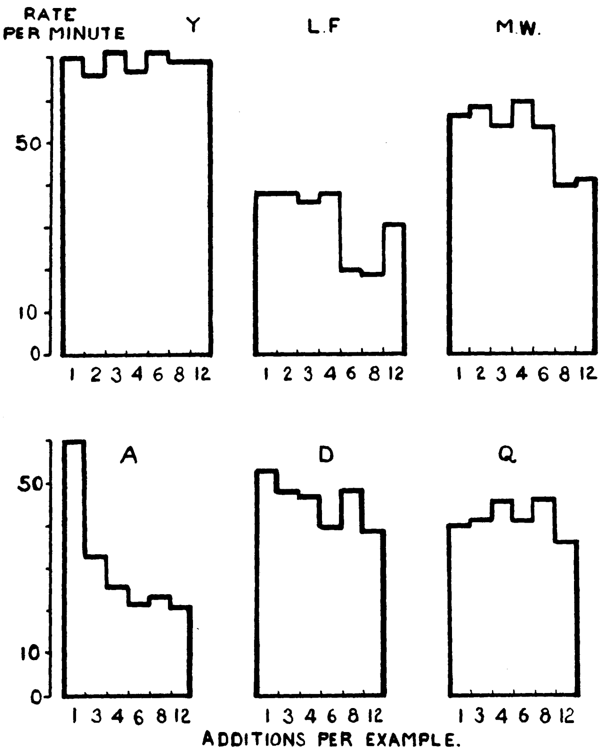

FIG. 13.—Change of rates with columns of varying length. Vertical scale = number of additions per minute; horizontal scale = number of additions in each example of the test. Test 1 corresponds to B in Table V, Test 2 to C, etc. Y, A, D, and Q represent the same individuals as in previous tables.

number of individual curves, so drawn as to make clear the changes in rates through the series. For curves Y, L, F, M, W the tests used were the same as those described above but the time allowance was just a minute for each test, and a longer rest was allowed between the tests. For during the second and third minutes of continuous work the scores even of adults begin to

fall off rapidly. Curves A, D, and Q are from the results given in the table. LF was seventh grade, and MW fifth grade. Individual Y (the writer) worked at a uniform rate, whatever the length of a column; the variations are of the order of chance variations. LF also worked a uniform rate for columns involving 1, 2, 3, and 4 additions, but an increase in the length of the column to 6 additions decreased the rate nearly one-half. In this case the cause is known; for the writer, watching, saw the girl when nearing the top of a column, hesitate, lose her place, and begin again. The limit of her attention span had been reached and it required an unusual effort, an extra grip on her attention, to reach the top of the column. Nearly every column was added twice with the consequent decrease in her rate. For MW much the same thing was true, but the drop came at 8 figures instead of 6 and the effect was not so marked.

In many cases causes were not as evident as in these two and all sorts of individual variations were found. A's curve has already been described. For minds of this type, the scores in the speed test give a totally false idea of the real ability in column addition. D represents a common type. Q illustrates the type that "warms up" to its work, but in all sooner or later a point is reached to go beyond which calls for greater effort, more inaccurate work, and rapid fatigue.

Out of 21 members of the eighth-grade class, the mistakes made were as follows:

Number of Test	Number of Individuals Making Errors	Number of Columns Added Incorrectly	Number of Test	Number of Individuals Making Errors	Number of Columns Added Incorrectly
Test A.....	2	2	Test E....	12	38
Test B.....	8	23	Test F.....	17	31
Test C.....	16	31	Test G....	21	61
Test D.....	17	44			

That is, here, as before, to make errors consumes time and the rate changes to correspond.

In bringing this discussion to a close, the writer feels that the evidence shows plainly that so far as chance variations are concerned, the result from a single measurement will ordinarily not

vary more than 5 points from the true score. That, on the other hand, the possibility of external accidents (breaking of pencil points, etc.) and of peculiar physical and mental condition radically modifying the achievement must *always* be kept in mind. That no attention should be paid to small variations from the standard, and that all large differences should be checked both by repeating the test and by comparison with similar scores in other tests; that the need for repeating measurements will occur at the most in about one case in ten. That if properly used, Test 1 measures one of the factors directly concerned in column addition and that the results rightly interpreted have a diagnostic value.

Finally, that the supreme thing in education is the fact of the very great variation in the abilities and needs of individuals. It is true that the writer urges the necessity for the measurement of the efficiency of the entire school, but it is also true that in no other way will the facts of the individual variation and of present gross inefficiency be revealed. The poorest school and the weakest teacher will, if they but keep at work enough years, turn out some naturally gifted individual to whose achievement they can forever afterward proudly point as a sample of their products. But at the present time no school has yet been found in which if the entire school be measured with standard tests under uniform conditions, the product of teaching is not shown to be so widely variable that so far as the particular ability is concerned, the school must be regarded as having failed utterly in accomplishing its purpose. At the present time the school is able to teach only those fitted by nature to respond readily to its teaching but if it were organized to detect and minister to the special need of the individual, vastly more could be accomplished. Definite aims, i.e., to render every child in the eighth grade able by June to add in four minutes 35 examples each a single column of 9 figures—and diagnostic tests—i.e., tests that will enable a teacher to determine exactly why a child can work but 23 such examples in the time allowed—and the experimental determination of efficient methods, are the lines along which progress will be made. The basic factor in education is thus the fact of individual differences in natural ability, and the supreme problem of future is the working out of administrative

methods of dealing with large masses of children, yet at the same time giving to each child the special attention and the special courses it needs, without sacrificing the benefits of class work and group instruction. Only as the Courtis Tests aid in this "discovery of the individual" and the laws of his development will they have accomplished their purpose.